Marie Sbrocca
10.10.2016
Perspectives on Computational Analysis
Paper #1

“But unless the public truly prefers a world in which nobody knows anything, more and better science is the best answer we have.” -- Duncan Watts [1]

We live in a brave new world of science, technology, and data. Over the past decades, our digital interactions have increased at an accelerated rate, and their nature has changed dramatically [2]. An increasingly robust amount of information about who we are, what we do, and who we know is recorded and analyzed.

Pop culture is quick to condemn new technologies and their potential vis-a-vis potentially destructive (though highly unlikely) paths they could take: as the beginnings of dystopia, or a future controlled by malevolent superintelligences [3, 4]. While this makes for engaging content, it reveals gross irresponsibility in journalism. Our societal perception is that science must either be composed of immutable truths, or it is likely false and untrustworthy [5,6]. Presenting new developments in dramatized, negative ways only reinforces these perceptions [7].

The outcry in both the public and academic spheres over the Taste, Ties, and Time study represents a particularly troubling instance of this kind of thought. While societal perceptions of science and technology are fraught with problems, academic outrage at addressable issues within new contexts is more surprising, and more dangerous [1]. Science and technology are generated on a continuum; they are imperfect arts, and new technologies are certain to have problems at their inception. Such problems should not immediately lead to public outcry or withdrawal of exploration, but rather to responsible, intelligent criticism, and re-evaluation of where risk is actually present.

Lewis et al. began their data collection in 2006, and their controversial paper was published in 2008. At this point in time, this type of research via social networking sites was relatively uncharted territory. As the authors point out, this represented the first data set of its kind to be made publicly available [8]. To appropriately evaluate the study, one must keep in mind the historical context in which its creators were operating: social media was not the near-ubiquitous layer over digital interaction it is today. Before evaluating the study according to the Menlo Report [9] and the ethical principles in *Bit by Bit* [10], it must be noted that guidelines were not published at the time the study was conducted; the guiding document available was the Belmont report [11], published in 1979 and non-inclusive of digital guidelines since it predates their need.

Thus, this becomes a retrospective analysis of a historical event according to a modernized set of principles. The time scale *is* objectively small, but given the pace at which technology progresses, this is not unreasonable [12].

Lewis et al. were mostly in compliance with the four principles of ethical digital research. We can reasonably assume that Harvard undergraduates are autonomous adult persons, who should be granted the opportunity to give consent (and that special considerations for protection are irrelevant to this population). Given that the researchers did not act in a way that mediated the subjects' experiences, but rather, studied extant Facebook data, violation of respect for persons seems a tenuous argument based on the semantics of what constitutes informed consent. Defining informed consent as a clause in a terms of service agreement may be contentious, but in this case, it is appropriate. Facebook's terms of service agreement is remarkably straightforward: in 2006, it was quite brief, and in 2016, it remains a reasonable length, and offers an interactive feature that explains data use [12, 13,14]. In both versions, it is stated that data, both user provided and inferred by interaction, may be provided to third parties, even for private profiles. The population at hand is educated, literate, and presumably intelligent; they have the wherewithal to either read the agreement, or to understand that they are signing something legally binding without reading it.

This also pertains to the researchers' adherence to the justice principle: the population studied was not exploited; they were capable of reading and understanding an agreement presented to them. While the study was completed on a clearly delineated demographic group (i.e. the Harvard class of 2009), it did not exclude any subgroup, and thus was acceptable.

The Menlo report explains beneficence as taking appropriate measures to do the most good, the least harm, and to manage risk. Given when this study occurred, the relative newness of such rich repositories of digital social information, and the lack of precedents for this type of dataset, the authors made a good faith effort to appropriately balance risks and benefits [8]. Based on the authors' discussion of privacy, it appears that they felt the benefits to social science community outweighed the potential risks to subjects. The problem is not disregard for risk, but a lack of realistic assessment of the probability of anonymization failure on the part of the researchers.

This study is not legally dubious. The researchers obtained permission from Facebook to use its data, which was collected in compliance with its terms of service. They were granted university permission to use university data, and were approved by their internal review board. In this case, the terms of service constitute an appropriate level of transparency, and a thorough explication of the research conducted could have resulted in observer paradox effects. Further, this principle seems inherently flawed as a heuristic, because it makes the egregious assumption that legal systems are agile enough keep up with the pace of technical innovation and societal progress [12].

The broader issue at hand is not compliance with the ethical research principles presented in the Menlo report; rather, it is that these principles are insufficient to induce appropriate action to protect and control digital privacy while not misapplying the precautionary principle at the expense of progress.

The response to this study suggests an ineffective and dangerous stance on data and privacy within the academic community: that is, we shouldn't conduct investigation or publish research because someone else might do something bad with that information. If information is accessible to researchers, it's accessible to hackers, and if it exists, someone is likely going to find a way to misuse it.

This mindset propagates two detrimental paths: first, we limit our own understanding of important topics, and second, we draw attention away from solutions such as developing better encryption, anonymization, and de-identification technologies, which seem more salient tools to keep private data private than not publishing it at all. We also conflate the issue of what actually needs to be private, and what is private because of outdated social constructs. For instance, social security numbers: this is the consequentially worst pieces of personal information that can be compromised. There is an undue focus on keeping these numbers private, when it is far more relevant to ask why we still use non-randomized nine digit numbers as the gold standard for identification, instead of developing better unique identifiers for people[1].

We retain ethical responsibility to do the right thing as a scientific community, but we also retain responsibility to act intelligently, and to optimize the efficacy of new technologies in advancing research, progress, and society.

To conclude, I would not use the dataset in my own research. I do not object to it on ethical grounds (though I would prefer to see another layer of anonymization prior to publication). Rather, the bigger objection I have to its use is relevance: it’s eight to ten years old. In this time, social media platforms have changed dramatically, as have our interactions with them, and the data that their purveyors collect [12]. In the interest of accurately representing interactions across social media, it would be more appropriate to obtain the most recent data available via the appropriate channels at Facebook.

[1] Note that as of June 25, 2011, social security numbers are at least partially randomized. [13]

Works Cited

[11] "The Belmont Report." HHS.gov. January 28, 2010. Accessed October 10, 2016. http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/#xbasic.

[14] "Facebook Terms of Service." Facebook. January 30, 2015. Accessed October 10, 2016. https://www.facebook.com/terms.

[15] "Facebook Data Policy." Facebook Accessed October 10, 2016. https://www.facebook.com/about/privacy/.

[13] "Facebook Privacy Policy." Internet Archive: Wayback Machine. October 23, 2006. Accessed October 10, 2016. https://web.archive.org/web/20061223233207/http://www.facebook.com/policy.php.

[7] "Mass Media News Coverage of Scientific and Technological Controversy: Edited Excerpts from a Symposium." Science, Technology, & Human Values 6, no. 36 (1981): 25-30. http://www.jstor.org/stable/689096.

[9] “The Menlo Report.” Center for Applied Internet Data Analysis. August 2012. Accessed October 10, 2016. http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf

[13] "Social Security Number Randomization." Social Security Administration. Web. Accessed October 10, 2016. https://www.ssa.gov/employer/randomization.html

[3] Achenbach, Joel. "Some Scientists Fear Superintelligent Machines Could Pose a Threat to Humanity." Washington Post. Accessed October 10, 2016. http://www.washingtonpost.com/sf/national/2015/12/27/aianxiety/.

[2] Gunelius, Susan. "The Data Explosion in 2014 Minute by Minute – Infographic." ACI. July 12, 2014. Accessed October 10, 2016. https://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/.

[6] Hopkins, Ryan. "Unbelievable: Why Americans Mistrust Science." Nature.com. March 13, 2014. Accessed October 10, 2016. http://www.nature.com/scitable/blog/scibytes/unbelievable.

[8] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. Social networks, 30(4), 330-342.

[5] Mooney, Chris, Illustration: Jonathon Rosen, Kevin Drum, Hannah Levintova and Tim Murphy, and Pema Levy. "The Science of Why We Don't Believe Science." Mother Jones. Accessed October 10, 2016. http://www.motherjones.com/politics/2011/03/denial-science-chris-mooney.

[10] Salganik, Matthew. Ethics, By Bit - 6. "6 Ethics." Bit By Bit -. Accessed October 10, 2016. http://www.bitbybitbook.com/en/ethics/.

[4] Shead, Sam. "Over a Third of People Think AI Poses a Threat to Humanity." Business Insider. March 11, 2016. Accessed October 10, 2016. http://www.businessinsider.com/over-a-third-of-people-think-ai-poses-a-threat-to-humanity-2016-3.

[12] Wadhwa, Vivek. "Laws and Ethics Can't Keep Pace with Technology." MIT Technology Review. April 15, 2014. Accessed October 10, 2016. https://www.technologyreview.com/s/526401/laws-and-ethics-cant-keep-pace-with-technology/.

[1] Watts, Duncan J. "Stop Complaining about the Facebook Study. It's a Golden Age for Research | Duncan J Watts." The Guardian. July 07, 2014. Accessed October 10, 2016. https://www.theguardian.com/commentisfree/2014/jul/07/facebook-study-science-experiment-research.